The Perils of Using Social Media Data to Predict the Spread of Diseases

Akshay Patel
University of Auckland
apat916@aucklanduni.ac.nz

Chintan Amrit
University of Amsterdam
c.amrit@uva.nl

David Sundaram
University of Auckland
d.sundaram@auckland.ac.nz

Abstract

The data produced by social media engagement is of interest to various organizations and has been used in different applications like marketing, finance and healthcare. Though the potential of mining this data is high, standard data mining processes do not address the peculiarities of social media data. In this paper, we explore the perils of using social media data in predicting the spread of an infectious disease; perils that are mostly related to data quality, textual analysis and location information. We synthesize findings from a literature review and a data mining exercise to develop an adapted data mining process. This process has been designed to minimize the effects of the perils identified and is thus more aligned with the requirements of predicting disease spread using social media data. The process should be useful to data miners and health institutions.

1. Introduction

Internet users use social media sites like Facebook and Twitter to share their life with others, including their health information. The potential to analyse this data has significant implications for infectious disease surveillance. Traditional flu surveillance performed by the Centres for Disease Control on Prevention (CDC) relies on outgoing patient information and test laboratory results. Such methods for infectious disease surveillance are outdated, as they confirm the outbreak of diseases two weeks after they have already emerged [1]. Big data-based analytics can predict the emergence of infectious diseases as they are spreading. By searching keywords such as "flu symptoms" in a region through social media data, the number of influenza patients in nearby hospitals can be predicted in advance. Resources like vaccinations or antibiotics can be distributed ahead of time to treat the number of expected patients.

The benefits of performing social media data analysis have been well researched [2]; conversely, the perils have been under-researched and we could only find a few popular publications [3]. The accuracy of predictive models is dependent on the quality of the input data and the transformation process used to produce meaningful results. Social media data is highly unstructured, creating a variety of troubles for data analysis, especially in the context of the spread of infectious diseases where an incorrect prediction could have a large undesirable impact [3]. This paper will help one understand the social, ethical, and technological implications of performing social media data analysis. Our findings will apply to social media data in general. This understanding will be used in the production of a context-adaptive approach to mining social media data, in the form of a data mining process. The lifespan of social media sites is uncertain, so the process should be generalizable to sites used at present and in the future.

The objective of this research paper, therefore, is to form an understanding of those problems and to find a way to mitigate them. Therefore, our research question is: *What are the perils of using social media for predicting the spread of communicable diseases?*

The communicable disease we focus on in this research is influenza and in particular we try to determine trend of influenza spread. Traditional influenza tracking by governments can take over two weeks to process, by which time the disease could have already spread to other people [1]. Data mining from Twitter can be used to minimize the time delay of information and get patients treated as soon as possible. In this paper, we analyse a Twitter dataset in order to predict if the users need an influenza shot. We then analyse this predictive model and describe the different perils of creating such a model by using the Knowledge Discovery in Databases (KDD) process. We then develop an enhanced data mining process that may help produce models with greater success rates. Medical institutions may also benefit from this research. If they have more accurate knowledge of the spread of disease, they will be able to more efficiently and effectively treat patients than they were before.

The rest of the paper is structured as follows; section 2 provides an overview of social media and

URI: https://hdl.handle.net/10125/59840
ISBN: 978-0-9981331-2-6

its perils, section 3 describes the research methodology we have used, section 4 the results and section 5 discusses our updated process model and section 6 concludes the paper.

2. Using social media data and its perils

Although social media data analysis shows lots of promise, there are many flaws and problems with it [3]. These problems are summarized against the KDD Processes in Figure 1 [4]. The perils are indicated in brackets with some perils (like 'Privacy and ethics') occurring in multiple stages of the process. KDD is a popular data mining process that is iterative (where previous steps may be revisited at any time during the data mining process). We now describe the perils listed in Figure 1 in more detail.

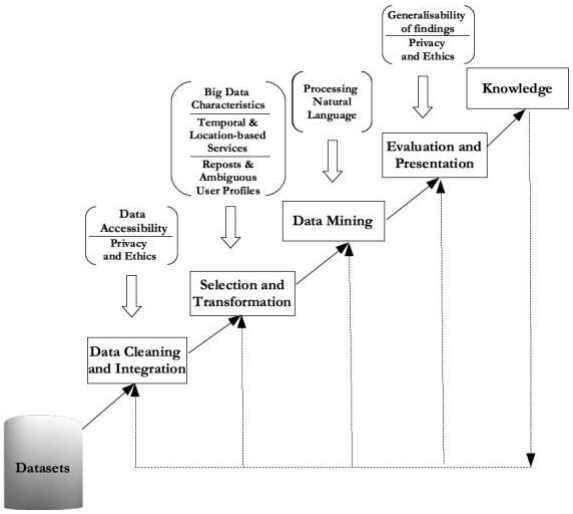

Figure 1: Perils of Using Social Media Data at the different stages of the KDD data mining process

2.1. Data accessibility

Anyone can access Twitter and see what people are talking about, but transforming that into valuable, actionable insights is tricky. To do so, one needs to have a firm grasp of the technical details of data collection and storage, statistical programming, predictive modelling and programming algorithms [16]. There is also the issue of having the technical capabilities to collect social media data [17]. There is a tremendous amount of initial and running costs associated with collecting, storing and developing algorithms to mine big data. The process is also time-consuming and complex [18]. Social media companies are also restricting access to their data. Facebook removed the ability for people to collect data from personal Facebook pages all together in 2015, over privacy concerns.

Twitter's developer platform provides two ways to access Twitter data APIs (Application Programming Interfaces): a standard version and an enterprise version. The standard version is free but provides only basic query functionality. The enterprise API is sold through resellers like GNIP (a social media API aggregation services), but at very high prices which few researchers would be able to afford [19].

There is a risk that because of the increased restrictions on social media data access, research in the area is becoming correspondingly confined. Researchers without access to the first-hand data will have to alter their research to match the data that just so happens to be available. Instead of the research questions and theories driving the data collection, the data drives what is being researched. This is a phenomenon known as "opportunistic data gathering" [17].

2.2. Big data characteristics

Content-based analytics is focused on user-generated data - such as text, voice, images and video. Such content is often high in volume, unstructured, noisy and dynamic. The characteristics of big data present challenges for analysing such content [5].

Social media data is generally unstructured [7]. Traditional data mining techniques were used to discover patterns and relationships from highly structured and relatively small datasets. Unstructured data does not fit into this traditional mould because it is so complex [8]. The Twitter API returns text-based datasets in JSON format. However, most data analysis disregards the images and videos shared through the platform, which could provide great insights.

The speed with which big data is created affects the speed at which we expect to analyse the data. As time passes, the value of social media data drops as the information becomes less current which has serious implications for data analysis. One of the important variables in monitoring disease spread is the location of the information. Lee and Kang [9] advise against storing geospatial big data in data warehouses and analysing them later. Instead, they suggest data miners should analyse this data on the fly and make decisions in real time. From a technological standpoint, the capability of fast accessing and mining big data is an obligation [5].

2.3. Temporal and Location-based services and triangulation

Despite its potential, Twitter location information is not well understood or well documented, limiting its utility [22]. The ability to triangulate the location of users is possible but flawed in many ways. Geo-location information provides the latitude and longitude information of users, this method for locating users is not very useful. Less than 0.5% of Twitter users enable the location function of their phones, mainly due to fears of privacy, bullying and stalking [23]. This creates problems with generalizability. There are many geo-inference studies where the self-reported location has been used instead of the GPS-based locations [24]. However, it yielded a much worse performance which leaves researchers with little choice but to use the location field which produces less generalizable findings [24].

2.4. Reposts and ambiguous user profiles

When analysing social media posts, one needs to be sure of the authenticity of the user and of the origin of the post. Some posts could simply be re-posts (re-Tweets) of other's posts and hence should be carefully considered in the analysis. Furthermore, analysis of Tweets is significantly affected by users submitting incorrect or outdated information which makes this factor less useful [15]. For example, a user may tweet about Australia, despite not living there or sending the tweet from that location. The content of messages may be indicative of a user's location, as they are more likely to speak about news related to their country of residence more than others. However, like the usage of the self-reported location in user profiles, this is subject to users speaking about locations where they do not reside [15]. Determining the location of users by studying their social networks is also flawed. It may work well on users who interact with a select group of individuals as they could be close friends. However, the more influential a user is on a social network, the greater the potential geographical diversity of their followers which affects the accuracy of this method [15].

2.5 Processing natural language

The ability of Web 2.0 technologies means that people from different countries and cultures can share information in a shared space. Language and culture vary by age, socioeconomic status, ethnicity, location etc. [10]. In the English language, a single word can have many different meanings. In isolation, a word could be interpreted in many ways. For example, the word "set" has 464 definitions according to the Oxford Dictionary. However, with context, the word can be globally understood. If a person asks "is everyone set?" it is commonly understood that the person is asking whether everyone is ready and/or prepared. Also, the meaning of words and phrases can change over time and refer to different things. For instance, the word 'wicked' used to mean something bad but in some current contexts young people use it to mean something wonderful! Thus, the current meaning of the terms needs to be considered when analysing the tweets. The trouble with text analytics is discovering that context. That will lead to finding the real meaning behind a social media post.

The characters from different languages work differently, which affects how different concepts can be discovered. In English, words are separated with space in between. However, in Mandarin, characters are not separated by spaces. According to a study by [11], 50% of tweets collected were in English. If we ignore social media data in languages other than English, we are effectively ignoring half the world.

Expressions of emotions are specific to language, culture and events. Quite often it can be found that the emotion behind a text cannot be explicitly read. Instead, the "triggers" of those emotions are written. The challenge is identifying terms that act as the trigger and associating them with an emotional label [13].

There are also massive negative implications to the character limit imposed by Twitter. The "brevity" of tweets requires users to include non-standard abbreviations, typos, irony, and trending topics called hashtags [13]. Such unconventional and unstructured texts are considered to be 'noise' as natural language processing (NLP) software does not handle such information so well, creating problems for Twitter content analysis [15].

2.6. Generalizability of findings

By default, Twitter activity is made public and therefore data is able to be collected by third parties. This is also true for things like blogs that do not require its users to have credentials in order to make contributions. Unlike Twitter, private Facebook page data is no longer available to API users. The discrepancy between how each social media site handles its data policies has the potential to introduce bias into data analysis [20]. Conclusions made from data analysis are highly dependent on the data used.

John et al. (2011) [21] found that there were personality differences between people who prefer Facebook or Twitter. They compared users of the two social networking sites based on the "Big Five" personality traits: neuroticism (emotional control); extraversion (sociability); openness (novelty-

seeking); agreeableness (friendliness) and conscientiousness (work ethic). It was found that Facebook users tended to use the site to mitigate loneliness more than Twitter users did. The results of the analysis by [21] showed that people who are more sociable used Facebook, while those who were seeking cognitive simulation used Twitter.

2.7. Privacy and ethics

Social media produces an immense amount of big data which contains highly interconnected information about its users. When those pieces of information about users are put together, the privacy of the individual is compromised [5]. In certain domains like social media, there is a fear that organizations will know too much about individuals. Most people have an awareness of privacy concerns, especially on social media and tend not to willingly provide confidential personal information online. Privacy has become a greater concern in the era of social media. To perform data analysis with accurate results, precise information may be needed. In data mining activities where the location of users is important, the latitude and longitude of users in real time are crucial. If users are afraid to provide this type of information, the data will be less useful, and the findings will not be as insightful as it could be.

3. The Data and its analysis

The data analysis method we follow is based on the one suggested by Shmueli and Koppius (2011). A dataset (flu dataset) with 7,062 tweets containing the word "flu" was collected between September 2015 and April 2016. This is a third-party dataset and is intended to capture the sentiment around the influenza virus [25]. The dataset was initially in the JSON file format and was initially read through Notepad. The data appeared to be reasonably clean and contained tweets in different languages but did not contain emojis. Almost every tweet contained the word "flu". The data was first exported to a Microsoft Excel spreadsheet and separated into columns using the built-in query function. Once the data was correctly separated into columns, it was much easier to assess the quality of the data. The data was visualized in Tableau from which some insights emerged.

A random sample of 400 tweets was collected using the random function ("RAND") of excel. A random number was assigned to the 7,062 tweets collected and were then sorted from lowest to highest. The first 400 rows were selected. In order to create our training set, each tweet was read by one of the authors. The tweets were then manually classified into two levels. The first level determined whether a tweet was relevant to needing a flu vaccination or not (0 = irrelevant and 1 implies it is relevant), despite the tweet containing the word "flu". The second level determined whether the content in relevant tweets indicated whether the person needed a flu vaccination or not (0 = user is unlikely to need a flu shot, while 1 implies that the user is likely to need a flu shot). We then used SPSS Modeller to select and filter the Tweets to include tweets relevant to influenza vaccination. After filtering out irrelevant tweets, we were left with 173 Tweets. Our purpose in analysing the flu dataset was to determine whether a Twitter user needs a flu shot or not. This requires classification of one of the variables. As there are only two possible values for the target variable, a binary classification is appropriate. We used a C&R (Classification and Regression) tree algorithm to predict the outcome (if the user needs a vaccination) in SPSS Modeler using 10-fold classification.

4. Results, discussion, and limitations

Using the Tweets, the model was able to determine whether the user needed the flu shot correctly 138 times and incorrectly 35 times leading to approximately 80% precision. The model could also be improved with enhanced textual analysis techniques like that done by Lamb, Paul, and Dredze [26], who built a classifier that was able to separate self-reported flu infections versus general flu awareness (e.g. "I think I am getting the flu and need to see a doctor" versus "John has the flu and needs to see a doctor"). We now discuss the different perils we found during our analysis using the terms introduced in Section 2 (also see Figure 1).

4.1. Big Data characteristics

4.1.1. Unstructured input data. The data was highly unstructured. When Twitter data is collected from an API, it is received in a JSON file format. The process of sorting the raw data into columns took a strenuous effort. Initially, a tab delimiter (e.g. a comma or semicolon), was used to isolate each column in the text. The problem is that once the flu dataset got to the content of the tweets, the tab delimiter did not work as intended. Some tweets contained the tab delimiter itself and the field was unintentionally split in two. Some tweets also contained no values (not even a null placeholder) so this splitting method was not successful. The query function in Microsoft Excel was used to split the columns, instead. Each column heading had to be carefully identified, split the

columns accordingly and finally check that the data matched that headings.

4.1.2. Input data quality. The Twitter data was collected using the keyword "*flu*", using Twitter's API. Most of the tweets were on topic and were speaking about the flu, its symptoms and other medical information. However, this was not always the case. Because social media users have the freedom to share (almost) any information they like, the data ends up being very noisy. The keyword used to search for tweets related to flu ended up returning results that were irrelevant. For example, one user tweeted "*does the flu shot also prevent Bieber Fever?*". Disentangling facts from perceptions and beliefs is quite important in such analysis.

4.2. Temporal and Location-based services and triangulation

The meaning of words and phrases can change over time and refer to different things. Regarding influenza, it can change in two different ways. An "antigenic drift" is where small genetic changes to the influenza virus cause it to change over time. The change is small but is still big enough for the body's immune system to not recognize it. A significant change to the influenza virus is referred to as "antigenic shift" which enables the flu virus to jump from one animal species to another. Most people have little to no protection against this type of virus [27].

In the flu dataset, the tweets were restricted to a period between 2015 and 2016. In 2015-2016, a Zika epidemic broke in the Americas. Twitter users may have used the word "flu" to refer to that specific disease. However, if the same search method was applied to Twitter data in a slightly different period, the word "flu" might have referred to another strain of influenza. For example, in November 2002 an outbreak of severe acute respiratory syndrome (SARS) occurred in China. Some of its symptoms are "flu-like" and include a fever, cough and sore throat. If social media were around during the 2002 SARS outbreak, users might have simply used the word "flu" to describe their symptoms before receiving an official diagnosis. Although both the Zika virus and SARS can be described as the flu, their treatments might be slightly different. This might mean that data mining processes should take temporal effects into account, as it affects how people communicate about phenomena. There is also a geographical effect as diseases in different locations can differ.

An analysis of the location of users according to the "country" variable revealed that most tweets have a "null" value which indicates that the user has turned their location-based services off for Twitter on their device. The top three values for the country field were as follows: Null - 6,834 records (96.77%); United States - 131 (1.85%); United Kingdom – 43 records (0.60%).

4.3. Reposts and ambiguous user profiles

On Twitter, the feature to re-post information is called "retweets". We found an example of this in the flu dataset. One user, "User 1" tweeted the following: "*RT @User2: I just have the flu*". Data miners may think that User 1 has the flu, when, they do not. The fact that users can re-post information creates complications as to who the information actually belongs to and can interfere with the accuracy of model building. The problem for data miners is that the information in the original post only truly belongs to the original uploader of the information. When users retweet something, the text in that post becomes attributed to the person who retweeted it.

4.4. Processing natural language

The software uses the bag-of-words (machine learning) model to retrieve information from text. The model illustrates that a piece of text (which could be a sentence) can be represented by a selection of words. Each word carries a certain connotation or mood (joyful, sad, positive, negative etc). Depending on the frequency of individual words (unigrams) or consecutive words (bigrams) in the text, the bag of words model classifies the document. The model was unable to correctly classify 100% of the tweets. In this data analysis, false negatives are where the user needs a flu shot but the model predicted that they do not need it. False positives are where a user has been determined to not need the flu shot. These are, of course, dependent on the manual labelling being correct and on the fit of the algorithm on the particular data. One of the common perils of classifying text from social media data is misclassifying a sentence based on unigrams and bigrams. For example, the Tweet "*flu shots hurt*" was classified as needing inoculation. Because the data analysis software uses a machine learning model, there is no human input regarding how the unigrams and bigrams classify the text. It is possible that the word "hurt" was determined to be a symptom of the influenza virus, and thus, the text was classified as the user still needing the flu shot. The results of the data analysis demonstrate how the context is important in data analysis and where the system can go wrong in terms of unsupervised learning.

4.5. Generalizability of findings

The flu dataset was only collected from Twitter; if a similar study using data from Facebook was performed, the results might have been different. This is a problem as only a segment of the population uses Twitter, and they might have different characteristics to the general population. Because Twitter gives open access to Twitter data, data miners prefer it over Facebook. The problem is that inferences made from the results of Twitter data may not be generalizable to Facebook users.

5. Adapted data mining process

The value of social media data changes rapidly over time. If a person makes a post that indicates that they have a virus, they obviously need treatment. This information is most useful as soon as the post is made. However, if the same text was analysed a week later, the value of that social media post is less useful. The user might have already recovered from their illness, or, passed it on to another person. There are other issues on these sites like the scarce availability of geospatial information and the use of language. What is required is an adapted data mining process that is contextualized for the peculiarities of this data mining task. The process needs to be specialized for text mining, unstructured data, geospatial data, and disease spread, all in one.

The proposed process will be based on the KDD data mining process by [4], but with modifications that effectively deal with the perils we mentioned in Section 2 (and Section 4). This process still retains the iterative aspects of KDD as it allows (and encourages) revisiting previous steps in the data mining process to optimize results. The data mining process is also intended to be generalizable to social media sites in general. The reason for this is that the lifespan of currently popular social media sites is uncertain, and if they were to suddenly drop in popularity, this data mining process would have little use. By preparing a data mining process that is generalizable, it could be used for social media sites that are popular in the future.

5.1. Step 1: Application domain

Health institutions are interested in predicting where diseases will spread in the future. Individual health institutions may only be interested in a disease which spreads occasionally or will use the model on an on-going basis to manage the seasonal flu. Depending on the goals of the health institution, this step will be revisited often. Some application domains might be more sensitive to noise in the data. If the institution is interested in monitoring the spread of a disease during an epidemic, the search results used to search for conversations on that disease in that period are likely to be on topic. On the other hand, if the institution is interested in the on-going spread of an infectious disease (e.g. the common flu), the search results might be more sensitive to noise in the data because people speak about diseases more casually on a normal basis.

Adaptation: Understand the application domain and identify the goals of the data mining process and perform a preliminary assessment of how noisy the dataset is likely to be. This is a standard sequence of tasks in this step and requires no change from the standard KDD data mining process.

5.2. Step 2: Data selection

The target dataset should be collected from social media sites using keywords identified in step 1 of this data mining process. Not only must the name of the disease be included, but related terms must also be included. For example, the term “flu” was used to search for data in 0, but it should be accompanied with other terms like “flu” or “sickness” and its symptoms such as “fever” and “cough”.

Users can post information about current events and other thoughts at the speed of a few clicks. We expect the data to be analysed just as fast [9]. In an organizational context, storing data and analysing it later is common practice. For example, analysing data about employees requires relatively little urgency. However, in the context of understanding disease spread and social media data, there is a greater need to analyse the data in almost real-time. Infectious diseases are a permanent issue and so real-time data collection is a necessity. By the time the data was manually labelled and processed, the user’s flu might have already spread to others. Individual health institutions will generally only be interested in diseases that are currently spreading in the area surrounding their location. If that is the case, the data collection method should focus on that place.

Adaptation: Predicting disease spread is an ongoing problem and the data collection method should reflect that. Instead of collecting historical data and analysing it later, what should be done is a live collection of data as it is matched with the search parameters. The data should be cleaned, analysed and visualized in real-time. There is a tool called PageLever which provides analytical tools for brands

that manage Facebook pages. The data can be collected in real time. Twitter provides two options for streaming live tweets which offer varying levels of capabilities: The POST statuses/filter API for standard users and the PowerTrack API for premium operators. Connection to the API is made by forming an HTTP request and consuming the stream for as long as is necessary. Only a single connection is made between the client machine and the API and new results are sent to the local machine. The data should be processed as it is collected, which can be costly in terms of having the infrastructure powerful enough to do that. One problem at this step that cannot be managed is Facebook's policy of not allowing third parties to analyse private Facebook pages. The decision has negative impacts for data analysis as it limits the amount of information about social media users that can be connected. Twitter, on the other hand, provides open access to users' data. Another problem is that data from multiple social media accounts cannot be collected for privacy reasons [5].

5.3. Step 3: Pre-processing and cleaning

Step 3a: Parsing. Social media data arrives in a highly unstructured format, such as a JSON file for historical data. The data is separated by columns and it is not ready for data analysis at that stage [7].

Adaptation: The data requires parsing before it can be read by the machine. Once the data has been sorted into variables then the variables can be worked with.

Step 3b: Filtering out shared information. Data that is "shared" on respective social media sites show as duplicate information when analysed. On Facebook, the function is named "share" and on Twitter, it is called a "retweet". Re-posted information can create complications for disease detection because one person who has a disease and requires treatment may post that information online. Other users can re-post that information using the retweet or share functions even though they do not necessarily experience the same feelings as the original poster. We found an example of this in the flu dataset. One user, "User 1" tweeted the following: "*RT @User2: I just have the flu !*" Data miners may think that User 1 has the flu when they do not. 24 out of 173 (14%) of the tweets were retweets, some of which were retweets of the same original tweet. This may have had a negative impact on the data mining results.

Adaptation: We want to filter out reposted information as the same text can show up repeatedly and does not reflect how individuals are feeling about their health [28].

Step 3c: Location triangulation. Earlier it was found that a very low proportion of users provided location information in their tweets from the flu dataset. Out of 7,062 individual tweets, 6,834 (97%) did not include location information. The single country which had the most tweets attributed to it was the United States, with only 131 tweets. These results could be attributed to privacy concerns [17]. The low adoption rate of location information sharing is a problem in the context of disease spread. The proportion of users who enable location information may not accurately represent their local population (including non-social media users). Thus, the data that is present may not be as useful as it could be.

Adaptation: One option is to completely disregard tweets which do not enable location-based services, as we cannot identify their exact location. This is not recommended as only a very low sample of social media users would be used to represent the general population. We propose a sub-step whereby the location of users is estimated. Instead of relying on users to turn on location-based services, a combination of their attributes should be used to triangulate their location. Variables could include their self-reported location in their respective social media profiles or the content of their social media posts. Another way in which the user location could be derived is through the language setting. This would be useful for predicting disease outbreaks where users do not enable their location-based services. This way a higher sample of the population can be understood. Of course, this is subject to some problems. The self-reported location is not accurate as users can report that they belong in erroneous locations like Mars. Message content is also variable as users can discuss locations they do not reside in.

Step 3d: Converting languages. Social media is used across the world. As a result, people from different cultures and regions may use those sites in different languages. The problem with this is that the different languages cannot be read using most semantic analysis technologies. For example, if some users use the social media site in Italian, an English-configured piece of software may not recognize it. It was found that most users spoke in English, but other users also spoke in Italian, French etc.

Adaptation: To overcome this problem, translation software should be integrated into the client machine. Some products can convert over 130 languages into English and will make it possible to analyse the text. This is not a perfect solution as the

text might not convert correctly and the meaning embedded within the text could be lost in the process.

Step 3e: Style of speech. Even if the language is converted correctly, there is still a problem of the exact same language and text being used in different contexts. The meaning of individual words and phrases can be different. A simple example is the use of the phrase "you alright?" In British culture, this is commonly interpreted as a greeting (like "how are you?). On the other hand, the phrase "you alright?' in an American setting could be interpreted as a concern for another person (e.g. if they were having difficulty in performing a task). During the data analysis stage of the data mining process, some tweets had the text that read "I'm dying from the flu". In some countries, this could be taken quite literally, in which case the person would need urgent medical care. In other countries, where the use of hyperbole is more prevalent, this information could be taken less seriously as the user is feeling ok. These contextual differences create problems in textual analysis.

Adaptation: The location of users could be used to manage the variation in the meaning of text depending on location. Because the languages have already been converted to English (above), there is less variability in the data. Depending on which location the user resides in, the "bag of words" used to analyse the textual data can be contextualized so that it accounts for different locations. So, the algorithm used to analyse text from British users is different from the one used for Americans, for example. Of course, this is subject to situations where people travel across countries but still maintain their previous style of speech.

5.4. Step 4: Data reduction and projection

Only useful features should be included in the dataset. Information such as the user's profile background colour is irrelevant when predicting disease spread.

Adaptation: No adaptation required. This is a standard process.

5.5. Step 5: Match goals to data mining method

What we are interested in is predicting the spread of infectious disease. The job of the data miner is to find a way to understand what people on social media are saying and determine where those users are moving. By blending the findings of the two together, only then can the spread of disease be determined. *Adaptation:* In this scenario, we are looking at the classification of text and the prediction of user location over time that is Time Series Classification and Prediction.

5.6. Step 6: Exploratory analysis and model and hypothesis selection

As identified in Step 5, this is a classification and prediction task. The problem here is that the data from social media is extremely volatile and there is a lack of control as to what users say and where they are located.

Adaptation: What is needed in this step are algorithms that are robust enough to handle social media data. The first stage is a text mining algorithm that classifies text. The bag of words approach should be able to first determine whether the social media post is relevant to an infectious disease. Second, the algorithm should determine if the user has the disease or if they had it in the past. If they had it in the past, then it is also irrelevant for the purposes of the data mining task. If it is determined that the user still has the disease, it would be of interest to the client. These are the classification elements of the process. Last, the location of the social media population should be monitored over time to predict where a disease is spreading. If a high population of users are speaking about a disease in a geographic location it is more likely that a disease will spread from that area to another if people are frequently travelling to and from that area. The results of these classification and prediction algorithms should be blended together to predict where diseases may be spreading. The results from this step are still subject to noise and volatility because of the nature of social media data.

5.7. Step 7: Data mining

Some of the problems with social media data like the triangulation of users and the language used have been addressed in the pre-processing stage of the data. However, as the data is being mined, new knowledge might be acquired that needs to be addressed. It is not just in the pre-processing stage where the values for fields can be determined. For example, in the pre-processing stage, the location of users was triangulated using information like their self-reported location. In the data mining stage, the data miner might find that the triangulated location of the user is incorrect as the user has indicated in their social media post that they belong to a different location. This would result in an incorrect location-profile being used for the user. This problem could also be applied to determining whether the user has a disease. For example, if a person says "I am so happy

after that concert. It was sick", the data mining algorithm might have determined that the person is not feeling well which alerts the health institution who could treat them. What the person means is that the concert was "great". The data miner should, therefore, adjust the bag of words used to analyse the textual data by revisiting previous steps in the data mining process.

Adaptation: Based on the new knowledge discovery from data mining, the data miner should revisit previous steps and adjust the algorithms to ensure the model correctly classifies values. This can be applied to determining the location of users and the classification of text in the social media posts. Although this is a function of the standard KDD process, it is especially important for social media data as it is of such a quality that revisions to the data algorithms are more likely.

5.8. Step 8: Interpretation/Evaluation

There are various problems with social media data that make it different from organizational data. Firstly, the data lacks structure. It can be in the form of text, emojis, images, videos etc. [7]. There are also problems with the language used on social media like what languages are used, the definition of words, the context in which they are spoken [10] and how the use of language differs between social media sites [10]. These problems affect the interpretation of data mining results.

Adaptation: At the interpretation stage, more care and attention to detail is needed in the scenario of trying to predict disease spread. As previously mentioned, the incorrect classification of data could mean life or death for patients in some circumstances. On the other hand, health institutions must minimize costs associated with storage and staff in preparation for disease outbreaks, so an over-estimation of how many people need medical care in a population is also not good. The data miners must ask themselves whether the data mining results accurately represent what is happening in the real world, keeping in mind that the text that is being analysed can be highly ambiguous. Not only must the data miners learn from the results of the data analysis itself, they must also keep in contact with the key stakeholders involved with the decision making from the data mining results (such as a high-level manager within a health institution). The results of the data mining process should also be compared to healthcare data provided by the CDC to evaluate its probability of success.

5.9. Step 9: Action

It is likely that the knowledge discovered from data mining would be incorporated into the healthcare system directly. They need to know the most up-to-date patient information as possible and that information would be lost if they had to wait for reports to be prepared by others.

Adaptation: No adaptation required. This is a standard process.

6. Conclusions

Despite the downward trend in deaths due to infectious diseases [29], they are still a threat and need to be monitored carefully. The ability to predict the spread of disease using social media data has great potential. However, social media data is very unusual and presents problems to data miners that needs addressing. To answer our research question, we first arrived at a framework of perils of analysing social media data to create a predictive model by reviewing the literature and then fit the lessons from our analysis into our framework. These helped identify the perils of predicting the spread of infectious disease using social media data, which were mostly to data quality, textual analysis and geospatial data.

The main contribution of this research is the development of an adapted data mining process, which has practical contributions to data science and healthcare. For data miners, the process can be used to predict the spread of disease in a more informed manner. If data miners are aware of the problems they might face with social media data ahead of time, they can be more prepared to address them and have more clarity regarding their work processes. The data mining process is intended to be generalizable across different social media sites, which makes it more useful in the future. For healthcare institutions, the data mining process will be indirectly useful in treating and preventing the spread of infectious diseases. To do so, they need the correct amount and type of resources ahead of time as they can perform efficiently and effectively. This requires an accurate forecast of future patient numbers, which the data mining process is intended to provide.

7. References

[1] C. W. Schmidt, 'Trending Now - Using Social Media to Predict and Track Disease Outbreaks', *Environ. Health Perspect.*, vol. 120, no. 1, pp. 30–33, 2012.

[2] H. Schoen, D. Gayo-Avello, P. Takis Metaxas, E. Mustafaraj, M. Strohmaier, and P. Gloor, 'The power of prediction with social media', *Internet Res.*, vol. 23, no. 5, pp. 528–543, 2013.

[3] A. Pettit, 'The Promises and Pitfalls of SMR. Prevailing discussions and the naked truth', *Mark. Res.*, vol. 23, no. 3, pp. 14–21, 2011.

[4] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 'From Data Mining to Knowledge Discovery in Databases', *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[5] D. Che, M. Safran, and Z. Peng, 'From Big Data to Big Data Mining: Challenges, Issues, and Opportunities', *18th Int. Conf. DASFAA*, pp. 1–15, 2013.

[6] B. Marr, *Big Data in Practise. How 45 Successful Companies Used BIG DATA Analytics to Delivers Extraordinary Results*. Wiley, 2016.

[7] D. Gil and I.-Y. Song, 'Modeling and Management of Big Data: Challenges and opportunities', *Futur. Gener. Comput. Syst.*, vol. 63, pp. 96–99, Oct. 2016.

[8] V. Diaconita, 'Processing unstructured documents and social media using big data techniques', *Econ. Res. Istraz.* , vol. 28, no. 1, pp. 981–993, 2015.

[9] J.-G. Lee and M. Kang, 'Geospatial Big Data: Challenges and Opportunities', *Big Data Res.*, vol. 2, no. 2, pp. 74–81, 2015.

[10] Y. Kim, H. Jidong, and S. Emery, 'Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection', *J. Med. Internet Res.*, vol. 18, no. 2, 2016.

[11] L. Hong, G. Convertino, and E. H. Chi, 'Language Matters in Twitter : A Large Scale Study', in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, no. 1, pp. 518–521.

[12] P. Song, a Shu, and a Zhou, 'A pointillism approach for natural language processing of social media', *arXiv Prepr. arXiv ...*, 2012.

[13] A. Balahur and G. Jacquet, 'Sentiment analysis meets social media - Challenges and solutions of the field in view of the current information sharing context', *Inf. Process. Manag.*, vol. 51, no. 4, pp. 428–432, 2015.

[14] A. Rosen, 'Giving you more characters to express yourself', *Twitter*, 2017. [Online]. Available: https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html. [Accessed: 28-Sep-2017].

[15] O. Ajao, J. Hong, and W. Liu, 'A survey of location inference techniques on Twitter', *J. Intell. Mater. Syst. Struct.*, vol. 26, no. 5, pp. 599–613, 2015.

[16] B. Marr, *Big Data in Practise. How 45 Successful Companies Used BIG DATA Analytics to Delivers Extraordinary Results*. Wiley, 2016.

[17] K. Weller, 'Accepting the challenges of social media research', *Online Inf. Rev.*, vol. 39, no. 3, pp. 281–289, 2015.

[18] B. Daniel, 'Big Data and analytics in higher education: Opportunities and challenges', *Br. J. Educ. Technol.*, vol. 46, no. 5, pp. 904–920, 2015.

[19] J. Garside, 'Twitter puts trillions of tweets up for sale to data miners', *The Guardian*, 2015. [Online]. Available: https://www.theguardian.com/technology/2015/mar/18/twitter-puts-trillions-tweets-for-sale-data-miners. [Accessed: 15-Aug-2017].

[20] W. W. Moe and D. A. Schweidel, 'Opportunities for Innovation in Social Media Analytics', *J. Prod. Innov. Manag.*, vol. 34, no. 5, pp. 697–702, 2017.

[21] H. John, R. Moss, B. Mark, and A. Lee, 'A tale of two sites:Twitter vs. Facebook and the personality predictors of social media usage', 2011.

[22] S. H. Burton, K. W. Tanner, C. G. Giraud-Carrier, J. H. West, and M. D. Barnes, '"Right Time, Right Place" Health Communication on Twitter: Value and Accuracy of Location Information', *J. Med. Internet Res.*, vol. 14, no. 6, p. e156, Nov. 2012.

[23] R. Li, S. Wang, and K. C.-C. Chang, 'Multiple location profiling for users and relationships from social network and content', *Proc. VLDB Endow.*, vol. 5, no. 11, pp. 1603–1614, 2012.

[24] D. Jurgens, T. Finnethy, J. McCorriston, Y. T. Xu, and D. Ruths, 'Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice', *9th Int. Conf. Weblogs Soc. Media*, pp. 1–10, 2015.

[25] J. Cai, 'Using Machine Learning to Analyze Twitter for Real Time Influenza Surveillance', *Medium2*, 2016. .

[26] A. Lamb, M. J. Paul, and M. Dredze, 'Separating fact from fear: Tracking flu infections on Twitter', *Proc. NAACL-HLT 2013*, pp. 789–795, 2013.

[27] Centers for Disease Control and Prevention, 'How the Flu Virus Can Change: "Drift" and "Shift"', *CDC*. .

[28] J.-A. Yang, M.-H. Tsou, C.-T. Jung, C. Allen, B. H. Spitzberg, J. M. Gawron, and S.-Y. Han, 'Social media analytics and research testbed (SMART): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages', *Big Data Soc.*, vol. 3, no. 1, p. 205395171665291, 2016.

[29] World Health Organisation, 'Global Health Estimates 2015: Deaths by Cause, Age, Sex, by Country and by Region'. World Health Organisation, Geneva, 2015.